

Sensitivity of the Maximum Parsimony Algorithm to Missing Data

Jack K. Horner
Science Applications International Corporation
P.O. Box 3827
Santa Fe, New Mexico 87501 USA
jhorner@cybermesa.com

BIOCOMP06

Abstract

A phylogenetic algorithm computes a tree of distance relationships on a set, S , of phylogenetic descriptions (which may not be complete), given a phylogenetic-description transformation function, D , defined on S . Maximum Parsimony (MP) is a widely used phylogenetic algorithm that computes the shortest phylogenetic tree that represents the tree distances on S determined by D . To date, the sensitivity of MP to missing/incomplete data has not been systematically investigated. Although a general characterization of this sensitivity is intractable, robust empirical characterizations for typical MP configurations are possible. Here, I present an analysis of the sensitivity of several commonly used tree robustness metrics to missing/incomplete data, for a widely used MP implementation and typical MP problem set-up, applied to randomized “mid-sized” missing-data sets. The results show a counterintuitive limitation of one of those robustness metrics.

Keywords: phylogenetics, cladistics, maximum parsimony

1.0 Introduction

We can define a *phylogenetic description* of a taxon, x , as an ordered pair, $\langle x, p \rangle$, where x is a taxon name, and p is mapping from the set $J = \{0, 1, 2, \dots, k\}$, where k is a non-negative integer (in phylogenetic parlance, J is called a set of *characteristics*, or *characters*), into a finite set of characters (called *character-states*). A set, S , of phylogenetic descriptions is called a *phylogenetic data set*. For any $\langle x, p \rangle \in S$ and $\langle y, q \rangle \in S$, let D be any member of the family, F , of iterated functions ([12]) that transform p into q . An iteration of D is called a *step*.

If no two members of S have the same image in S under a member of F , then F determines a family of trees, G , defined on S . If we further map the phylogenetic distance between any two adjacent nodes in a member, T , of G to the link/edge between those nodes, then T is a weighted tree ([5]), and is called a *phylogenetic tree* or *cladogram*. The *phylogenetic distance* from x to y in T is defined as length of the path in T from $\langle x,p \rangle$ to $\langle y,q \rangle$ in T .

If O is a function (typically, some optimization function) that selects a member of G , the pair $\langle O, D \rangle$ is called a *phylogenetic algorithm*. *Maximum Parsimony* (MP) is a phylogenetic algorithm that determines the set of shortest trees (there might be more than one tree with the same shortest length), given S and D . A common variant of MP uses a heuristic, instead of an exhaustive, search for the smallest tree(s) to help mitigate the combinatorial growth of computational time. Typically, heuristic-search MP uses a random-number sequence to help mitigate sampling bias, and this gives rise to a distribution of trees.

A mapping, I , from $\langle O, D \rangle$ into a *biological* theory (typically of inheritance or evolution) is called a *biological interpretation* of that algorithm ([11], p.20). (A phylogenetic algorithm need not have a biological interpretation ([6], [9], [10]), but biologists are almost exclusively interested in those that do.)

There are several commonly used measures of the robustness of T . Among these measures are the *Consistency Index* (CI; [7]), the *Homoplasy Index* (HI; [8]), and the *Retention Index* (RI; [7]). Let

M_M be the maximum steps required to generate T

N_C be the number of characters in T

N_M be the number of characters in S

CI is defined as

$$N_M/N_C.$$

Homoplasy is any similarity found in two taxa which is not due to common transformation. HI is defined as

$$1 - CI.$$

RI is defined as

$$(M_M - N_C) / (M_M - N_M).$$

S may not be complete (see, for example, ([2]). To date, the question of how sensitive MP trees are to missing data has not been systematically investigated. A general characterization of this sensitivity is intractable because the relevant parameter space has ~100 dimensions, many of which individually have the power of the continuum. In actual practice, however, only a few of these parameters are typically varied, making at least some empirical characterizations possible.

Phylogenetic Analysis Using Parsimony (PAUP; [1]) is a widely used phylogenetic analysis software package that implements several phylogenetic algorithms, including MP. Here, I present an empirical analysis of the sensitivity of TL, CI, HI, and RI to missing/incomplete data, for PAUP's MP implementation, on a typical MP problem set-up, applied to a randomized "mid-sized" data set.

2.0 Method

All phylogenetic computations described in this section were performed on a 3.1 GHz Linux/Intel platform with 2 GB memory. A PAUP-compatible NEXUS file ([1]), B, describing 100 nominal taxa ("T1", "T2", ..., "T100") was generated using the *genbasedata* software ([3]). In B, each phylogenetic description consisted of 100 characters whose state-values were randomly drawn from a four-member character-state range ("0", "1", "2", or "3").

```
P A U P *
Portable version 4.0b10 for Unix
Sun Dec 18 00:32:56 2005
[non-essential text deleted here]
Heuristic search settings:
  Optimality criterion = parsimony
  Character-status summary:
    Of 100 total characters:
      All characters are of type 'unord'
      All characters have equal weight
      All characters are parsimony-informative
  Starting tree(s) obtained via stepwise addition
  Addition sequence: random
    Number of replicates = 10
    Starting seed = 1649913265
  Number of trees held at each step during stepwise addition = 1
  Branch-swapping algorithm: tree-bisection-reconnection (TBR)
  Steepest descent option not in effect
  'MaxTrees' setting = 1000 (will not be increased)
  Branches collapsed (creating polytomies) if maximum branch length is
zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted
```

Figure 1. PAUP/MP setup used in this study.

P percent ($P = [0 | 10 | 30 | 50 | 70]$) of the character-state values in the taxon-descriptors in B were randomly replaced with the PAUP “missing data” character (“?”) using the *gendatablock* software ([3]).

The resulting NEXUS files (1000 per missing data percentage) were analyzed using the maximum parsimony (MP) algorithm as implemented in the PAUP software, configured as shown in Figure 1. TL, CI, HI, and RI values were extracted from the PAUP log file for each run using the *getmetrics* software ([3]). This experiment was repeated 1000 times for each value of P, defaulting the PAUP random number seed in each PAUP invocation.

The sample mean and standard deviation of TL, CI, HI, and RI data obtained from the previous step were then computed using the *SIMSTAT for Windows* ([4]) software and graphed under the Microsoft Excel graph function. The results are described in Section 3.0.

3.0 Results

The entire set of phylogenetic calculations required ~200 CPU hours on the platform described in Section 2.0. Standard deviations of less than 0.005 were rounded to 0.00.

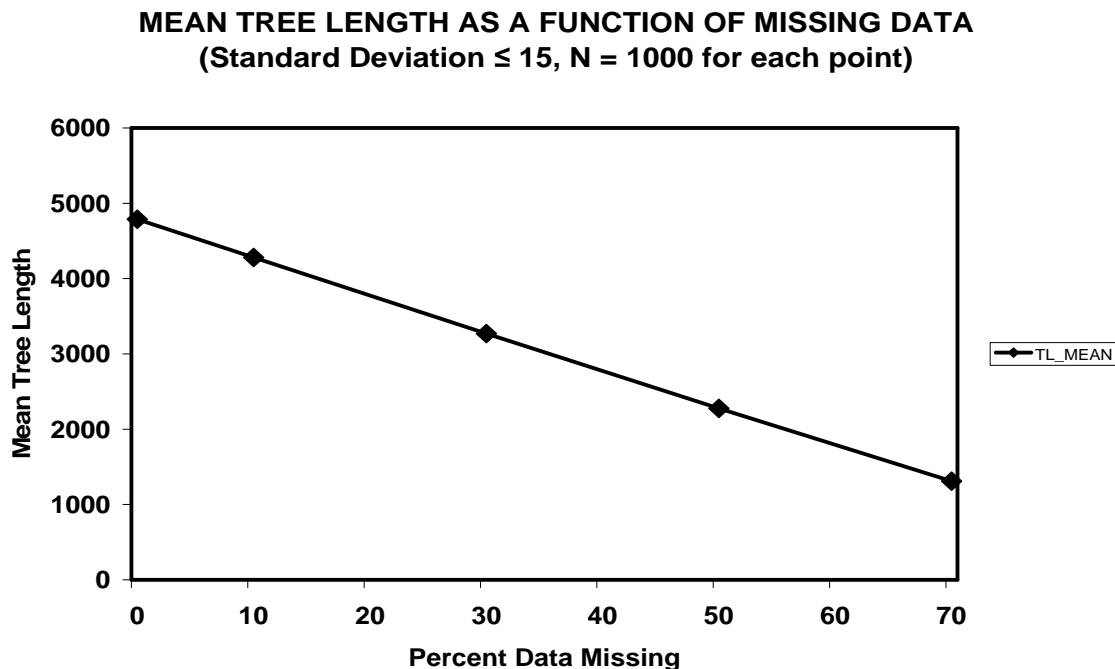


Figure 2. Mean tree length vs. percent missing data.

MEAN CONFIDENCE INDEX AS A FUNCTION OF MISSING DATA
(Standard Deviation = 0.00, N = 1000 for each point)

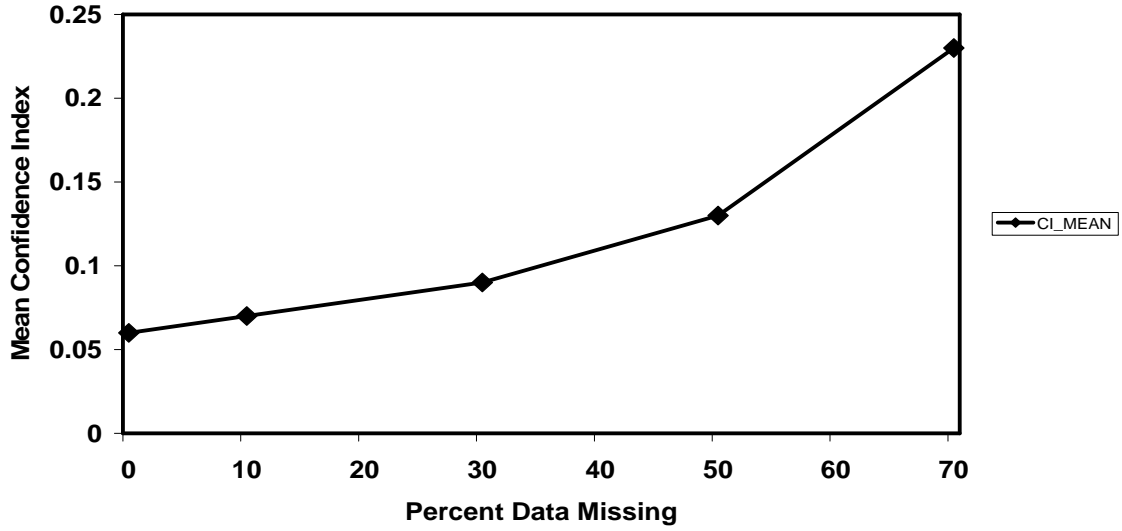


Figure 3. Mean Confidence Index vs. percent missing data.

MEAN HOMOPLASY INDEX AS A FUNCTION OF MISSING DATA
(Standard Deviation = 0.00, N = 1000 for each point)

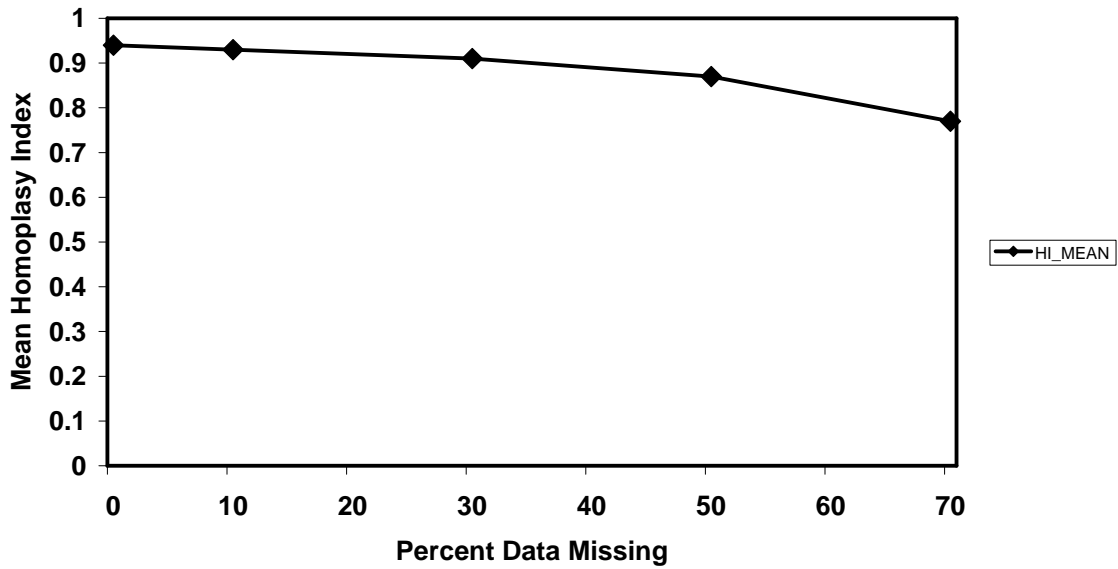


Figure 4. Mean Homoplasmy Index vs. percent missing data.

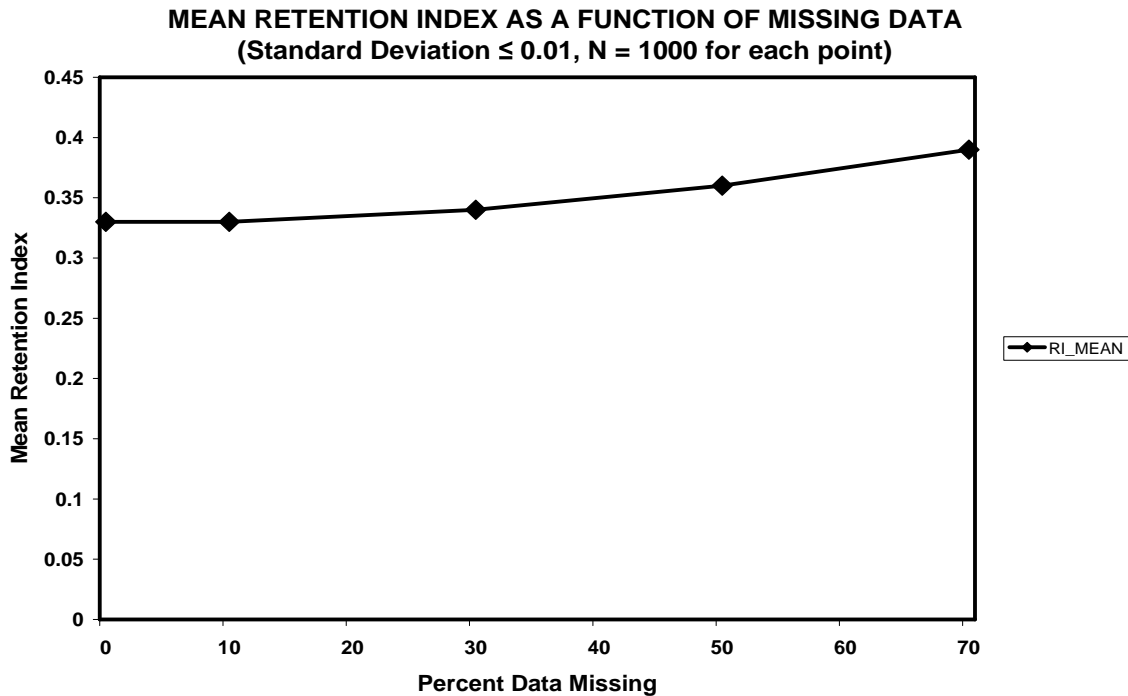


Figure 5. Mean Retention Index vs. percent missing data.

4.0 Discussion and conclusions

The results presented in Section 3.0 show that for the MP implementation, setup, and randomized data sets analyzed, mean tree length, which is a coarse measure of the discrimination among taxa in phylogenetic trees, decreases linearly in percent data missing. (The specific trees produced, of course, are likely to be sensitive to which specific character-values are missing.)

Mean CI counter-intuitively increases with percent data missing (Figure 3), revealing a limitation of this metric as a measure of “correctness”.

Mean HI (Figure 4) and mean RI (Figure 5) change ~15% across the range of conditions studied, showing that, for the conditions analyzed, MP’s detection of homoplasy is robust in the presence of missing data.

The results under setups different from those described in Section 2.0, of course, could differ from those described in Section 3.0. It would be informative, in any case, to apply the method described in Section 2.0 to phylogenetic data sets containing $10^3 - 10^6$ randomized characters per taxon. Statistics for the lower end of this range could be calculated in a month on five ~3 GHz workstations running concurrently; calculating the statistics for the upper end of the range in a month would require at least a loosely coupled network ([13]) of several thousand processors.

5.0 Acknowledgements

This research benefited from discussions with Bill Spangenberg and Jim Holten. For any errors or limitations, I am solely responsible.

6.0 Disclaimer

This work is not claimed to represent the views of Science Applications International Corporation or its customers.

7.0 References

- [1] D. Swofford. Phylogenetic Analysis Using Parsimony (PAUP). Portable version for Unix v4.0b10. URL <http://paup.csit.fsu.edu/>.
- [2] P. J. Makovicky, S. Apesteguía, and F. L. Agnolin. The earliest dromaeosaurid theropod from South America. *Nature* 437 (13 October 2005), pp. 1007-1011.
- [3] J. K. Horner. *genbasedata*, *gendatablock*, and *getmetrics* are perl scripts available on request from jhorner@cybermesa.com.
- [4] N. Péladau. SIMSTAT v2.09. Provalis Research. URL <http://www.provalisresearch.com>. 1996.
- [5] R. Diestel. Graph Theory. Springer. 1997.
- [6] J. K. Horner. A Neighbor Joining method for identifying a Stage I ovarian cancer signature in the mass-spectrum of serum proteins. Proceedings of the 2004 International Conference on Mathematical and Engineering Techniques in Medicine and the Biological Sciences. pp. 118-123.
- [7] J. S. Farris. The Retention Index and the Rescaled Consistency Index. *Cladistics* 5 (1989). pp. 417-419.
- [8] T. Sang. New measurements of distribution of homoplasy and reliability of parsimonious cladograms. *Taxon* 44 (1995), pp. 77-82.
- [9] M. J. O'Brien and R. L. Lyman. Cladistics and Archaeology. University of Utah. 2003.
- [10] P. M. W. Robinson and R. J. O'Hara. Cladistic analysis of an Old Norse manuscript tradition. *Research in Humanities Computing* 4(1996), pp. 115-137.
- [11] C. C. Chang and H. J. Keisler. Model Theory. North Holland. 1990.
- [12] A. F. Beardon. Iteration of Rational Functions. Springer-Verlag. 1991.
- [13] J. L. Hennessy and D. A. Patterson. Computer Architecture: A Quantitative Approach. 2nd Edition. Morgan Kaufmann. 1996. See especially Chapter 7.